Distilling Data from Large Language Models:
An Application to Research Productivity Measurement

Maya Durvasula, Stanford University
with Sabri Eyuboglu and David Ritzwoller

December 2024
mdurvas@stanford.edu
Data available at: https://github.com/DavidRitzwoller/pubmed_clinical_trials

This Project

Objective: Collect and organize the universe of publicly-available information on the design and statistical outcomes of (pharmaceutical) clinical trials

Publicly-available sources

- Scientific publications
- Regulatory approval documents (from e.g., FDA)
- Administrative database records (e.g., ClinicalTrials.gov)

Approach: extract structured data from unstructured text w/ LLMs

Challenge: value of data lies in our ability to control true/false positive rates

This Paper

We construct new data on the universe of published clinical trials indexed in PubMed / MEDLINE (2010-2022)

Primary contributions

1. A method & workflow for use of LLMs that captures the benefits of frontier, proprietary models
   - at a fraction (~3%) of the cost
   - with the transparency and reproducibility of open-source models
2. New data on the universe of clinical trials that
   - correct classification errors in existing data, which generate spurious findings of increasing clinical trial production
   - shed light on compositional changes in scientific publications relevant to measures of research productivity

Constructing Data on the Universe of Clinical Trials Disclosed in PubMed/MEDLINE

Sample of Interest

Prospective interventional clinical studies that primarily evaluate the effects of investigational or approved drugs on exclusively human subjects

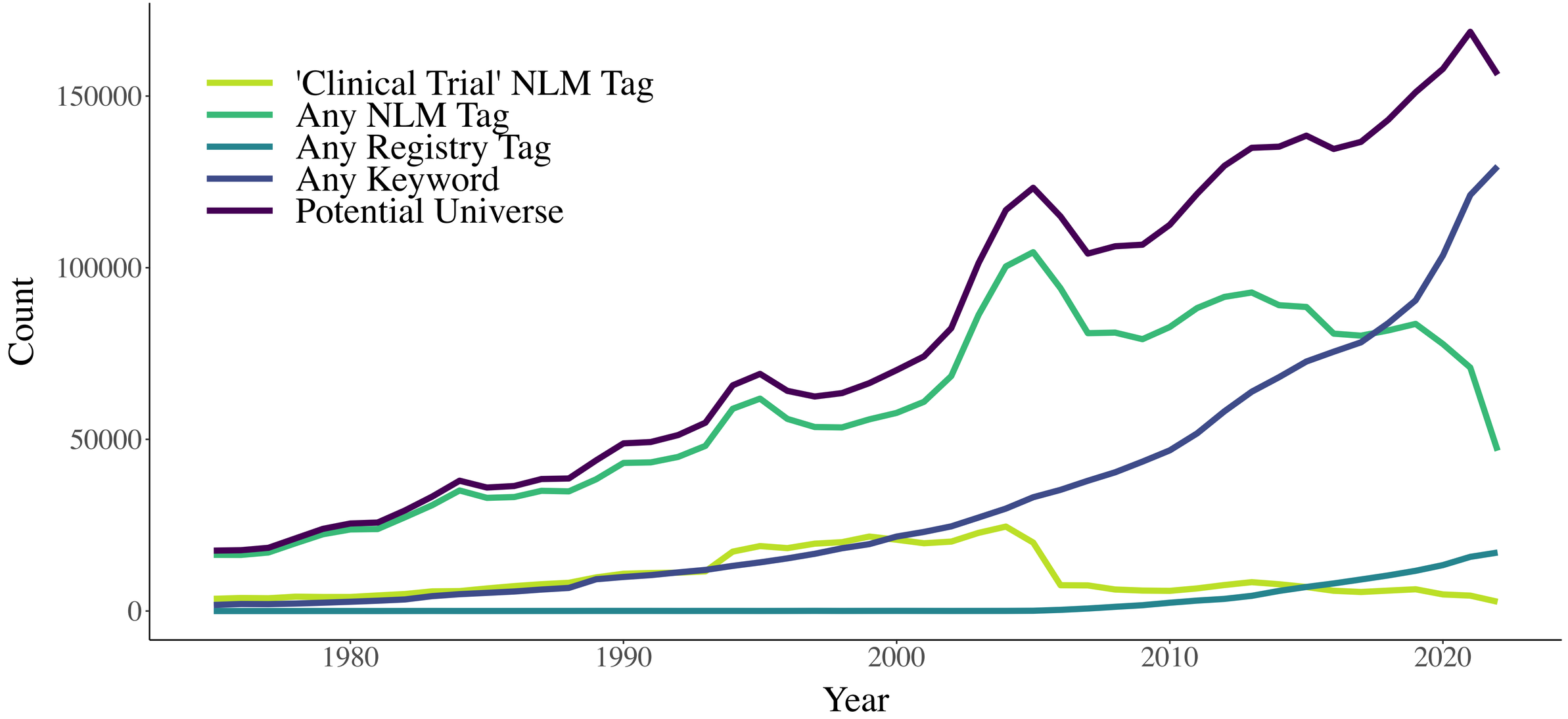
Publication occurs on or after 1 January 2010

Exclude if:

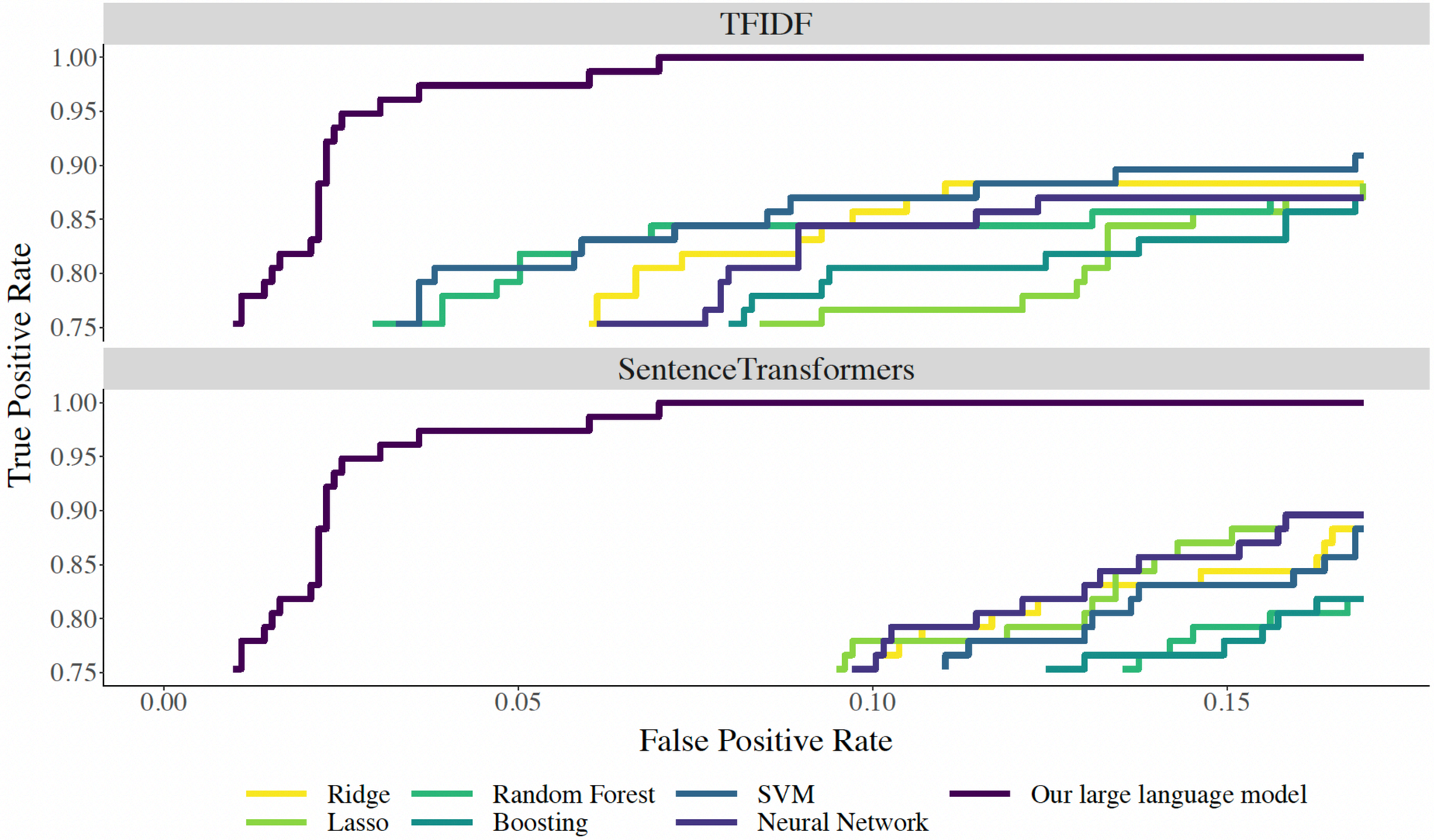
- Clinical trial study protocol
- Meta-analysis
- Observational study
- Dietary supplement, dietary choices, behavioral interventions, medical devices

Data: PubMed / MEDLINE: ~ 34 million records

Clinical Trials in PubMed / MEDLINE



Standard Machine Learning Algorithms



Constructing Task-Specific Language Models

- We construct a large language model optimized for our task, using **model distillation**, in four stages:

    1. Hand-labeling [~3k labels]
    2. Prompt Engineering [~3 types / 3 subtypes, paper details our error analysis]
    3. Noisy Label Extraction From Proprietary Models [OpenAI's GPT-3.5, GPT-4]
    4. Fine-Tuning [Set of open-source models]

[on **model distillation**—for construction of lightweight chatbots, see Taori et al. 2023, Chiang et al. 2023, Xu et al. 2023; for completion tasks, see Liu and Low, 2023; for API queries, see Patil et al. 2023]

At the end of this prompt, you will be shown an abstract from an academic publication indexed in the PubMed/MEDLINE database.

Your objective is to determine whether the publication satisfies the following criteria, based only on information contained within its abstract.

Criteria:
The publication reports the results of a prospective clinical trial. The clinical trial may be of any phase. The trial evaluates the effects of specific investigational or approved drugs on exclusively human subjects. The abstract is written in English.

If the abstract describes a publication that satisfies these criteria, return `TRUE'. If the publication does not satisfy all criteria, return `FALSE'. Do not return any extraneous text. You must return either `TRUE' or `FALSE'.

The abstract that you will consider is as follows:

Abstract: {abstract}

Answer:

Clinical Trial > N Engl J Med. 2020 Dec 31;383(27):2603-2615.
doi: 10.1056/NEJMoa2034577. Epub 2020 Dec 10.

Safety and Efficacy of the BNT162b2 mRNA Covid-19 Vaccine

Fernando P Polack [1], Stephen J Thomas [1], Nicholas Kitchin [1], Judith Absalon [1], Alejandra Gurtman [1], Stephen Lockhart [1], John L Perez [1], Gonzalo Pérez Marc [1], Edson D Moreira [1], Cristiano Zerbini [1], Ruth Bailey [1], Kena A Swanson [1], Satrajit Roychoudhury [1], Kenneth Koury [1], Ping Li [1], Warren V Kalina [1], David Cooper [1], Robert W Frenck Jr [1], Laura L Hammitt [1], Özlem Türeci [1], Haylene Nell [1], Axel Schaefer [1], Serhat Ünal [1], Dina B Tresnan [1], Susan Mather [1], Philip R Dormitzer [1], Uğur Şahin [1], Kathrin U Jansen [1], William C Gruber [1]; C4591001 Clinical Trial Group

Collaborators, Affiliations + expand

PMID: 33301246 PMCID: PMC7745181 DOI: 10.1056/NEJMoa2034577

Abstract

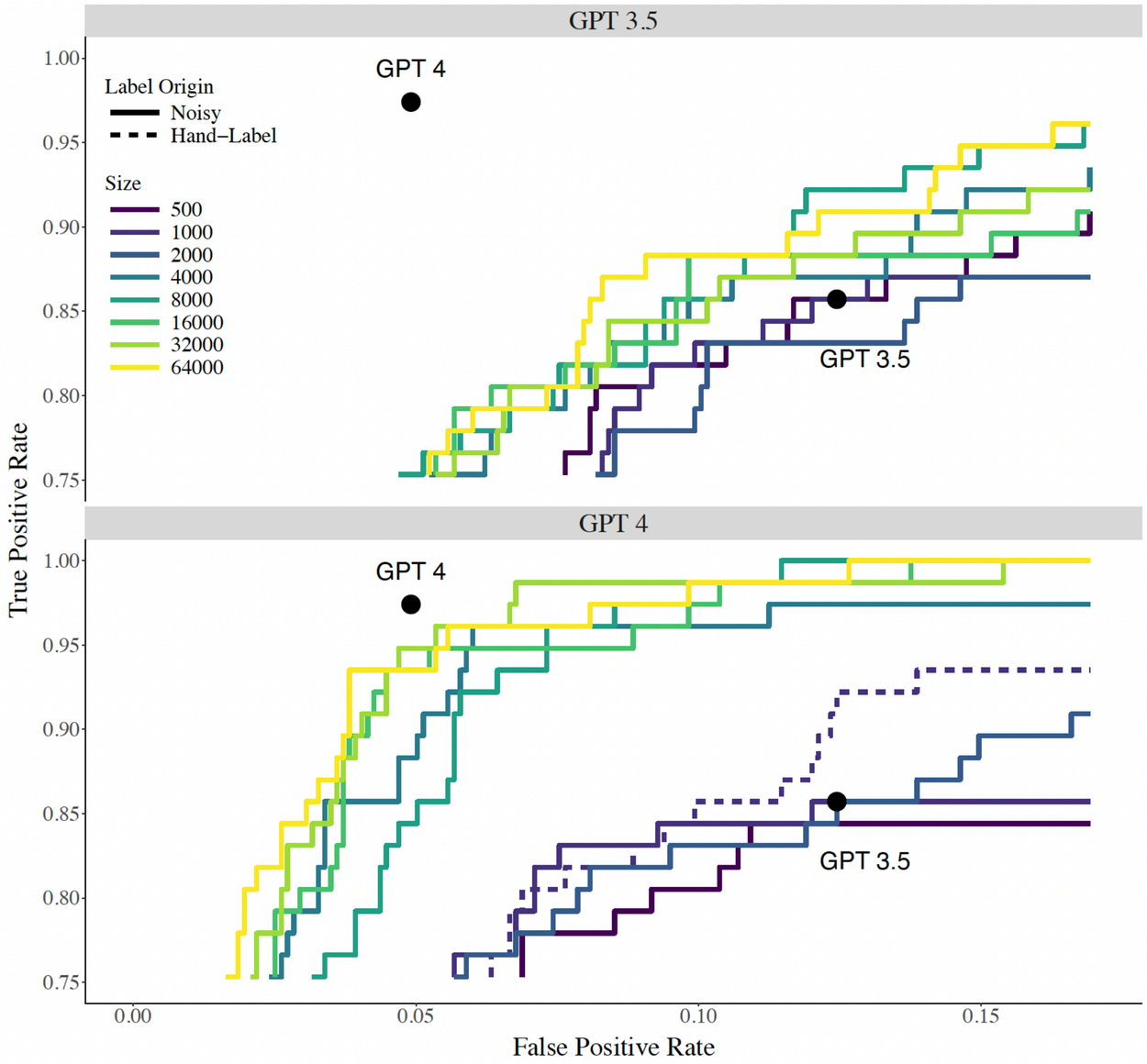
Background: Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infection and the resulting coronavirus disease 2019 (Covid-19) have afflicted tens of millions of people in a worldwide pandemic. Safe and effective vaccines are needed urgently.

Methods: In an ongoing multinational, placebo-controlled, observer-blinded, pivotal efficacy trial, we randomly assigned persons 16 years of age or older in a 1:1 ratio to receive two doses, 21 days apart, of either placebo or the BNT162b2 vaccine candidate (30 µg per dose). BNT162b2 is a lipid nanoparticle-formulated, nucleoside-modified RNA vaccine that encodes a prefusion stabilized, membrane-anchored SARS-CoV-2 full-length spike protein. The primary end points were efficacy of the vaccine against laboratory-confirmed Covid-19 and safety.

Results: A total of 43,548 participants underwent randomization, of whom 43,448 received injections: 21,720 with BNT162b2 and 21,728 with placebo. There were 8 cases of Covid-19 with onset at least 7 days after the second dose among participants assigned to receive BNT162b2 and 162 cases among those assigned to placebo; BNT162b2 was 95% effective in preventing Covid-19 (95% credible interval, 90.3 to 97.6). Similar vaccine efficacy (generally 90 to 100%) was observed across subgroups defined by age, sex, race, ethnicity, baseline body-mass index, and the presence of coexisting conditions. Among 10 cases of severe Covid-19 with onset after the first dose, 9 occurred in placebo recipients and 1 in a BNT162b2 recipient. The safety profile of BNT162b2 was characterized by short-term, mild-to-moderate pain at the injection site, fatigue, and headache. The incidence of serious adverse events was low and was similar in the vaccine and placebo groups.

Conclusions: A two-dose regimen of BNT162b2 conferred 95% protection against Covid-19 in persons 16 years of age or older. Safety over a median of 2 months was similar to that of other viral vaccines. (Funded by BioNTech and Pfizer; ClinicalTrials.gov number, NCT04368728.).

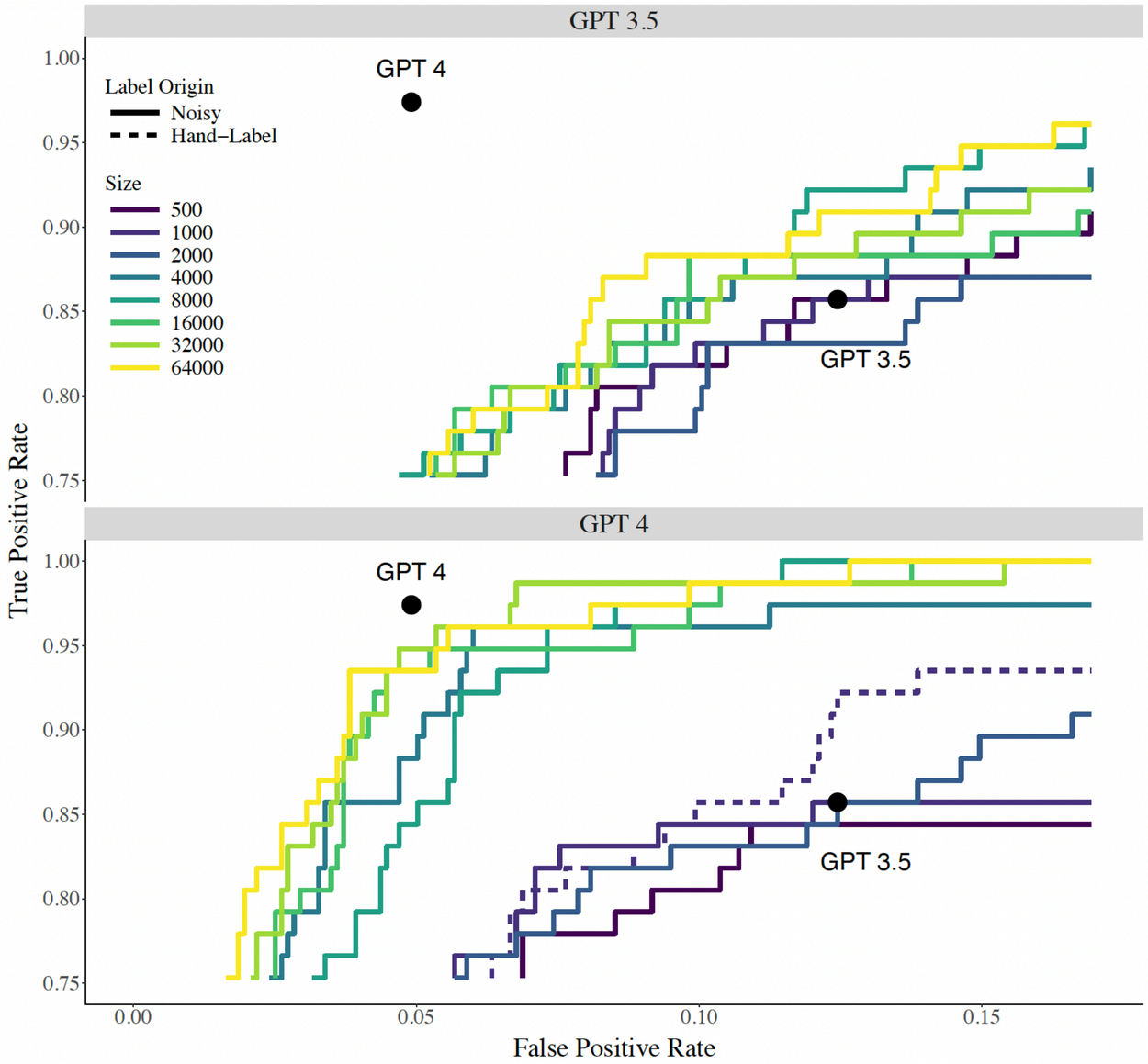
Constructing Task-Specific Language Models



- Compute 64k “noisy” labels for randomly selected publications using the best-performing prompts for GPT 3.5 and GPT-4.
- We use noisy labels to train off-the-shelf BERT models from two classes:
  - BigBird (125M + 355M param.)
  - BioMedBERT (125M + 355M param.)

[Comparable performance for 7B and 70B LLaMA, but much more complex to train]

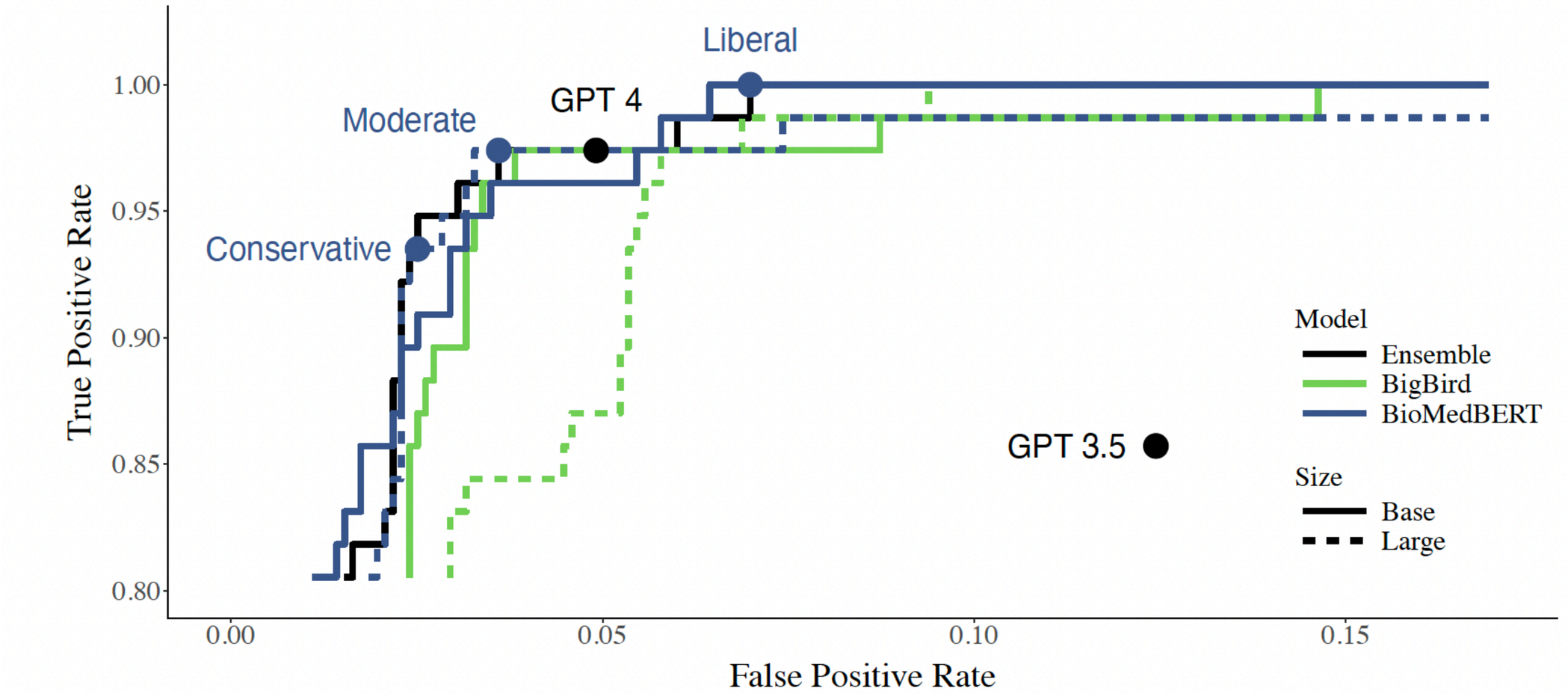
Constructing Task-Specific Language Models



Three observations

- We document a phase transition in model quality, in the number of training labels used (~8000 labels).
- Models fine-tuned with labels extracted from a noisier model (GPT 3.5) *exceed* the performance of GPT 3.5.
- Models fine-tuned with labels extracted from GPT 4 *match* the performance of GPT 4.

Constructing Task-Specific Language Models

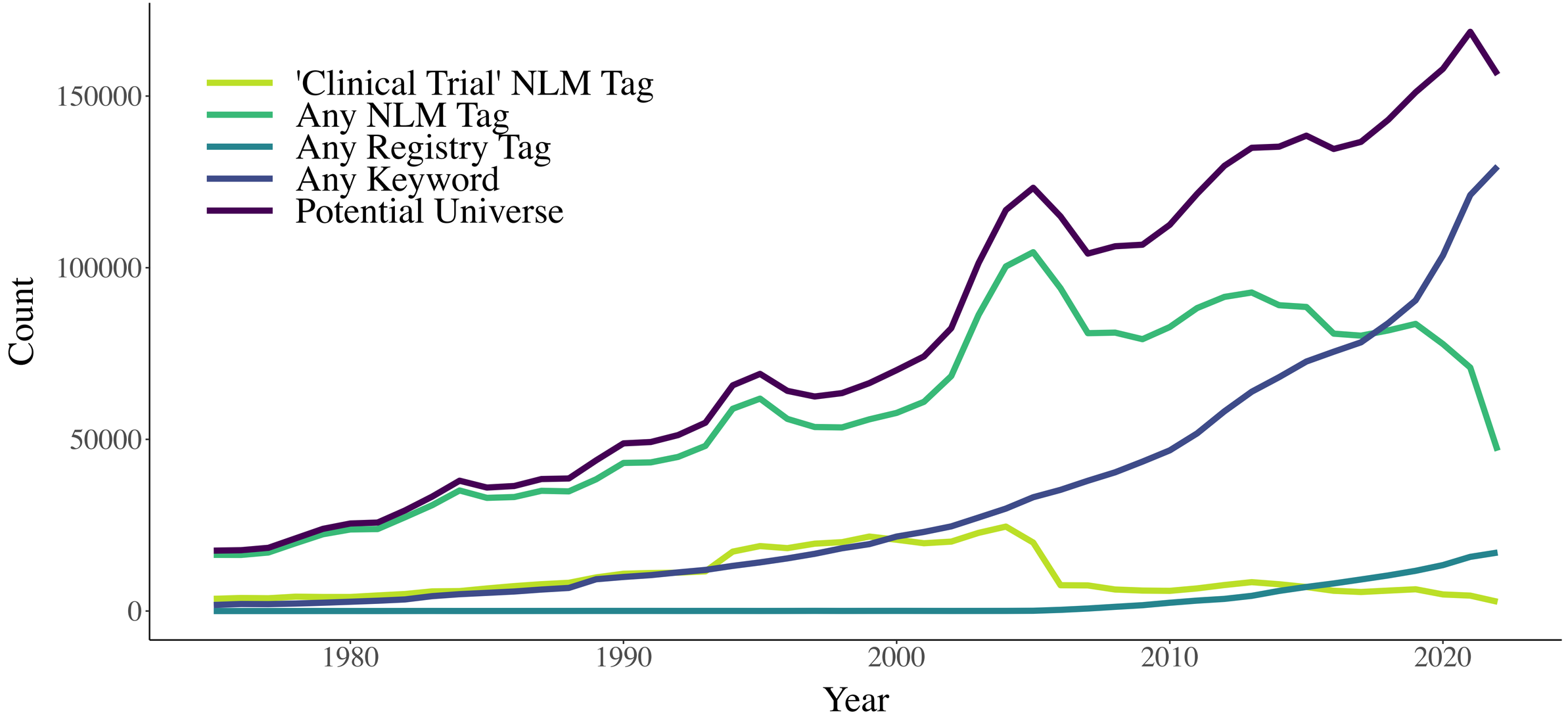


Preferred sample: **Conservative**

152,027 publications

Trends in Clinical Trial Production

Existing methods indicate sharply increasing trends …



cited as evidence for declining productivity in …

SCIENCE

Science Is Getting Less Bang for Its Buck

Despite vast increases in the time and money spent on research, progress is barely keeping pace with the past. What went wrong?

By Patrick Collison and Michael Nielsen

WORK IN PROGRESS

America Is Running on Fumes

In film, science, and the economy, the U.S. has fallen out of love with the hard work of ushering new ideas into the world.

By Derek Thompson

FUTURE PERFECT SCIENCE

Why is science slowing down?

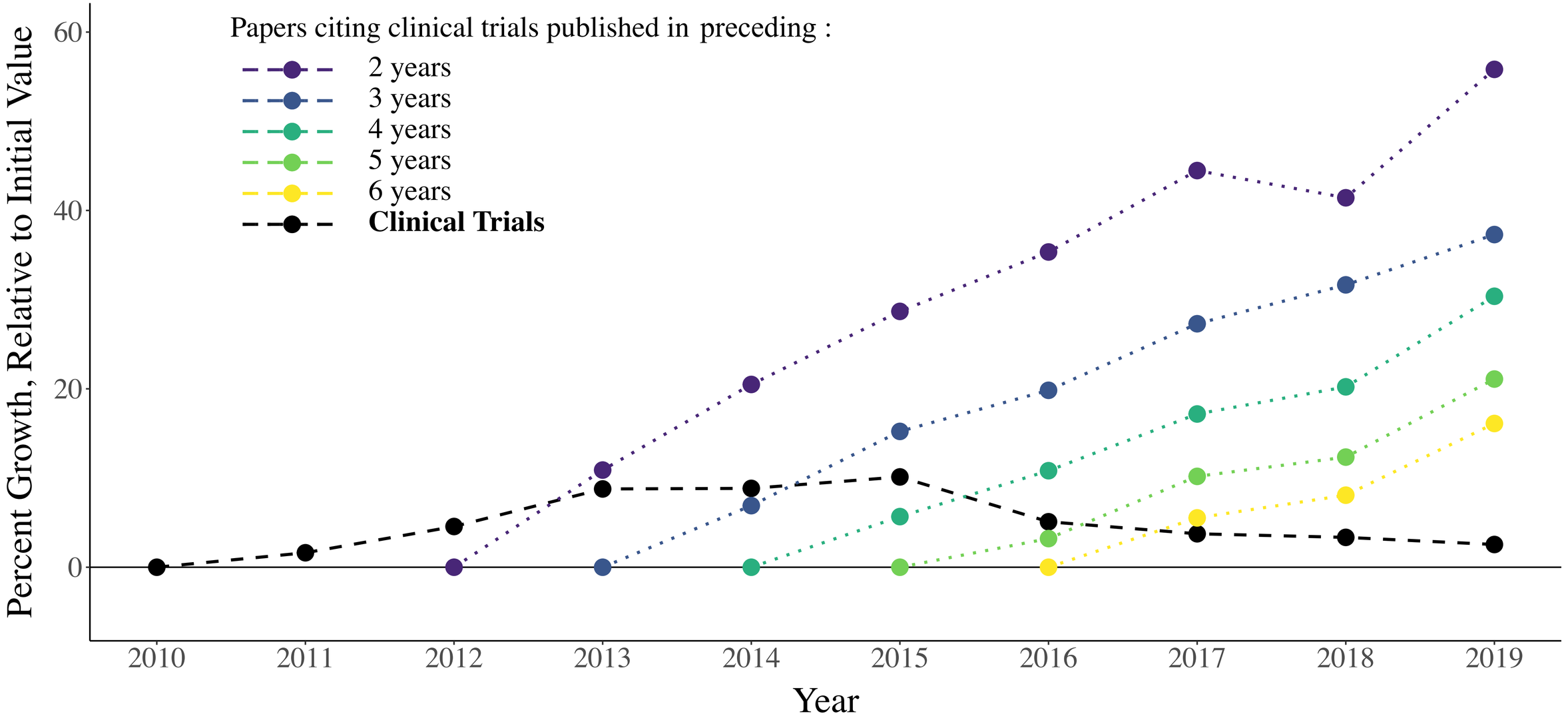
Science is the engine of society, and the decline of truly disruptive research is a warning sign for all of us.

By Kelsey Piper | Jan 11, 2023, 10:00am EST

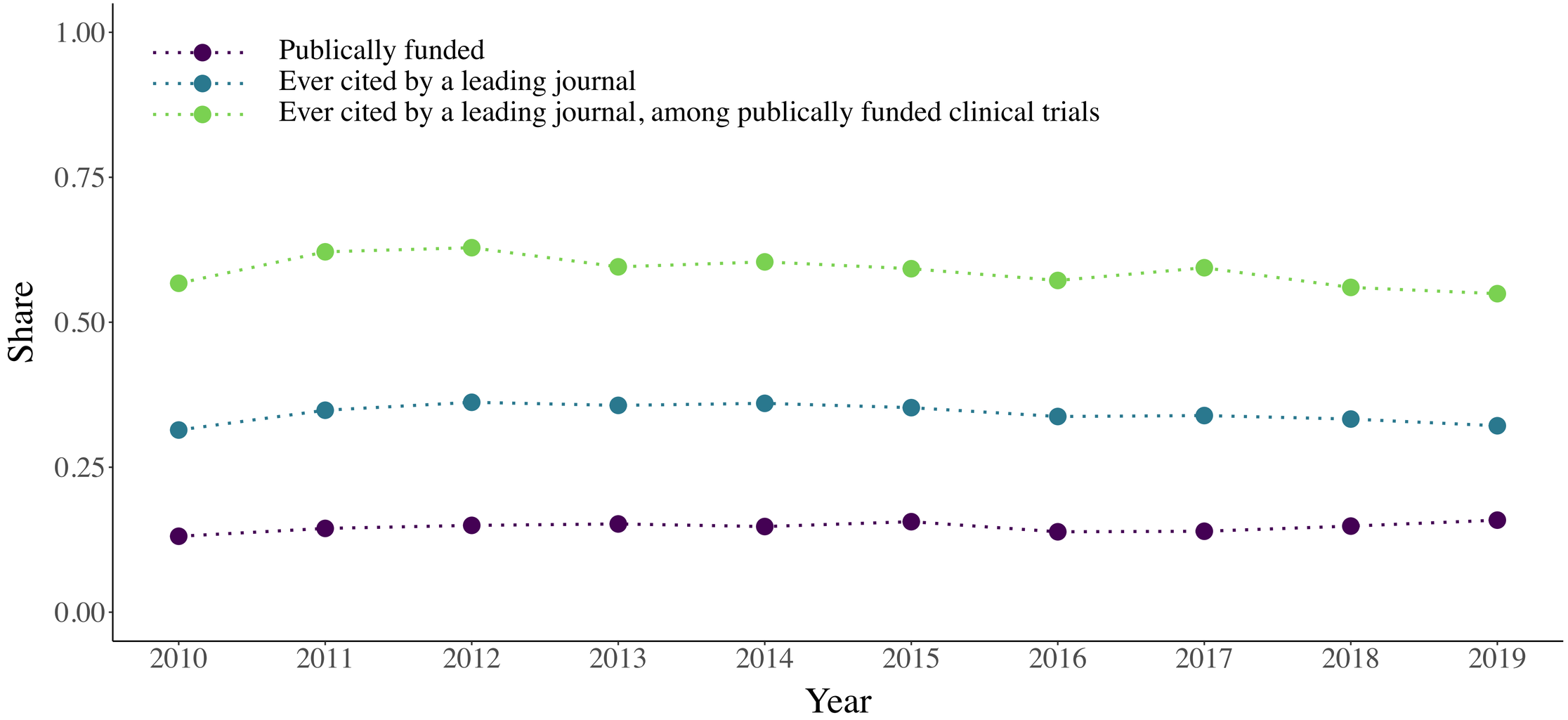
The Age of Decadence

[see Bloom et al. 2020, Goldin et al. 2024, Scannell et al. 2012, Pammolli et al. 2011, Ruffolo 2006, Cockburn 2004, 2006]

We find stability in trial *quantity*,



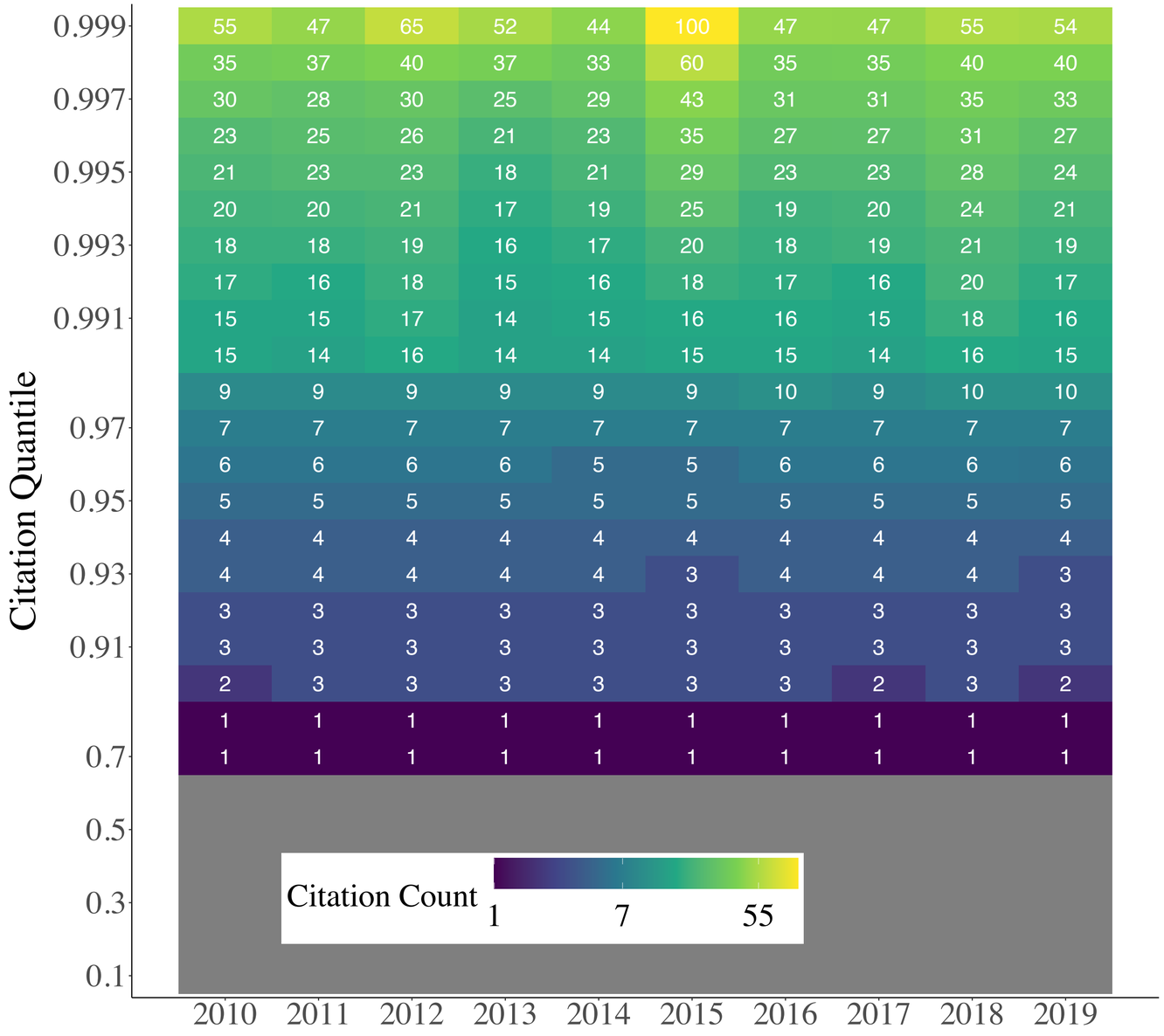
Trial *quality*



Composition

~60 percent of trials are never cited by a top-100 journal in medicine.

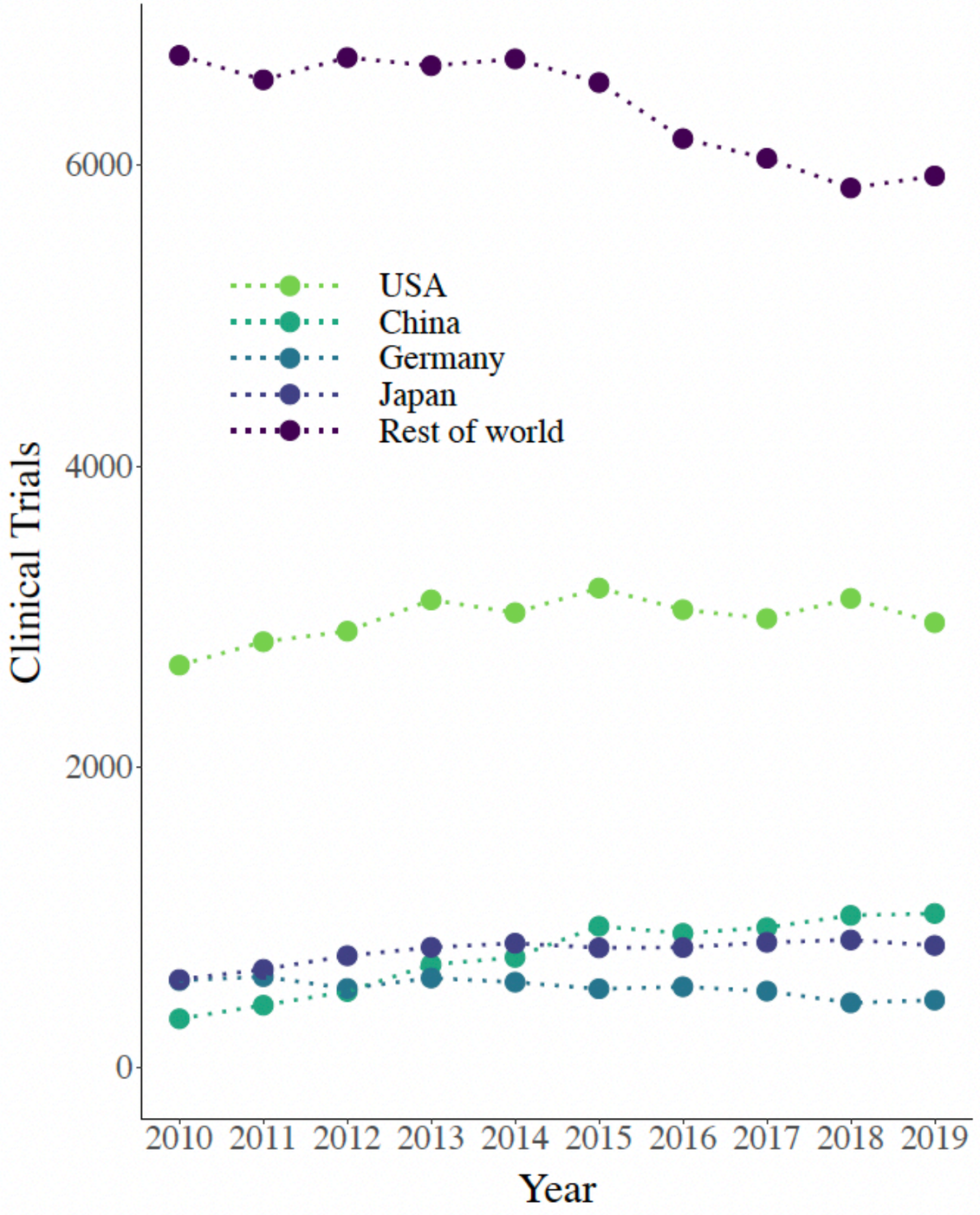
The “best” and “worst” papers receive the same number of (3-yr) citations in each period.



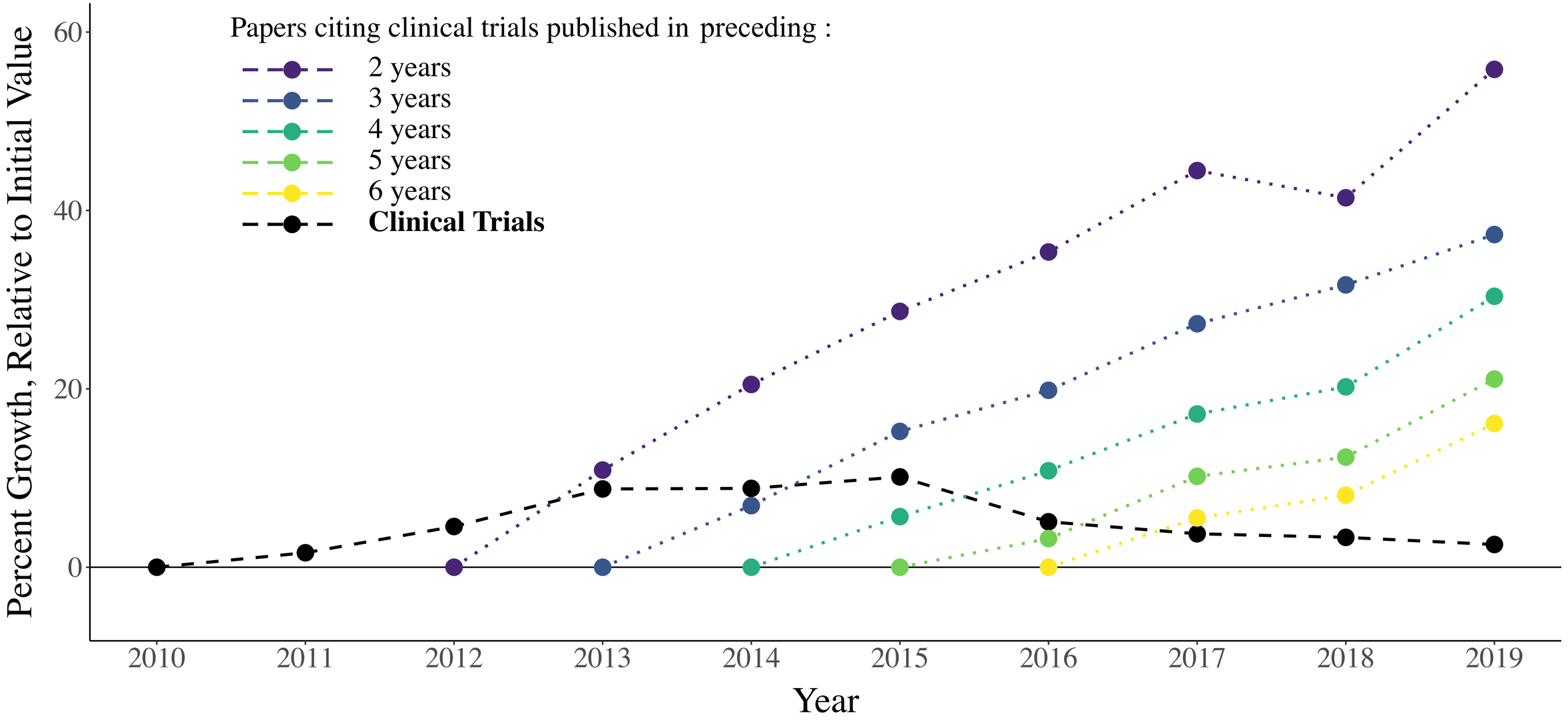
Geography

Top four producers:

1. United States
2. China
3. Germany
4. Japan

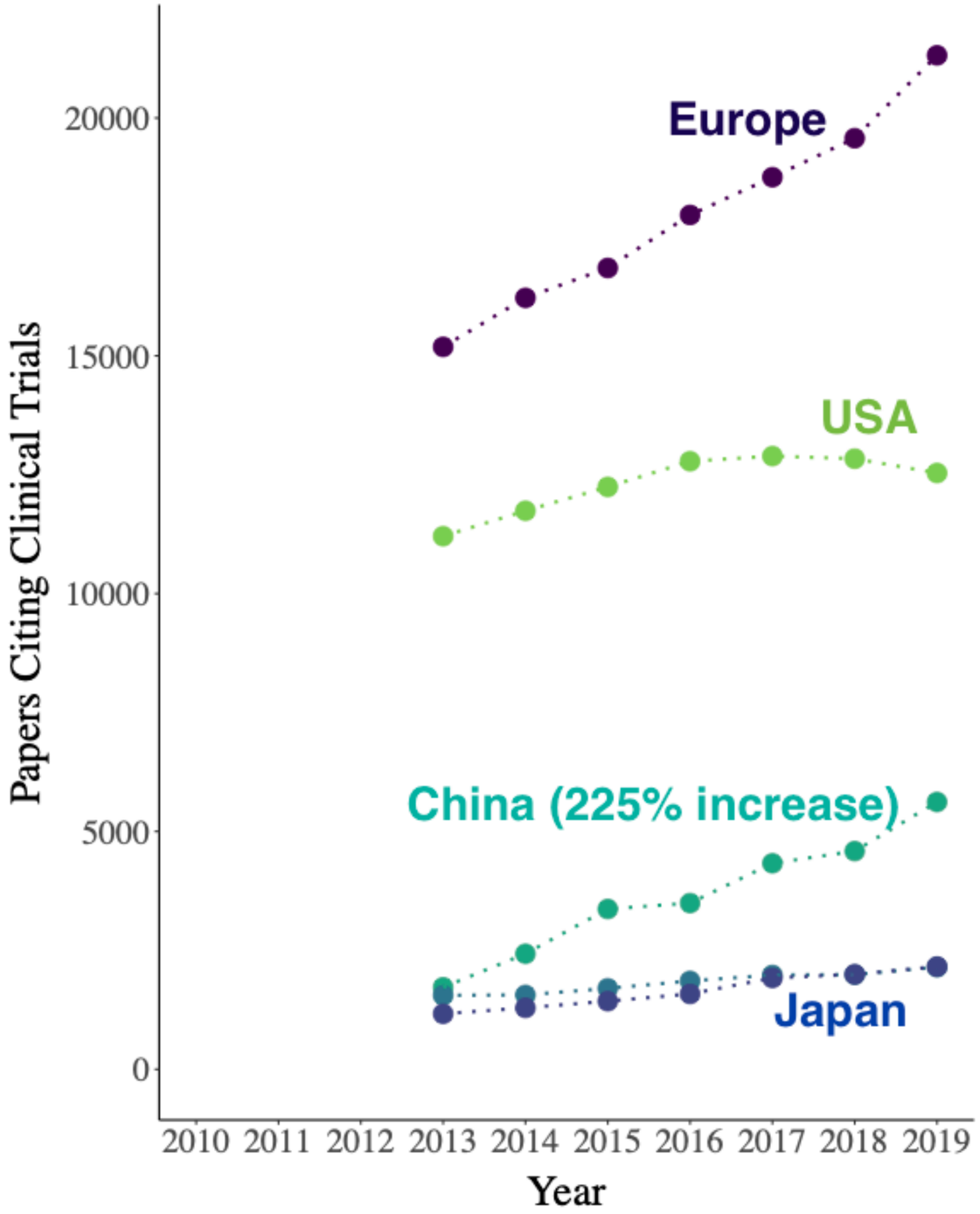


Classification errors capture growth in *textually similar*, non-trial papers



Growth explained by changes in *geography*

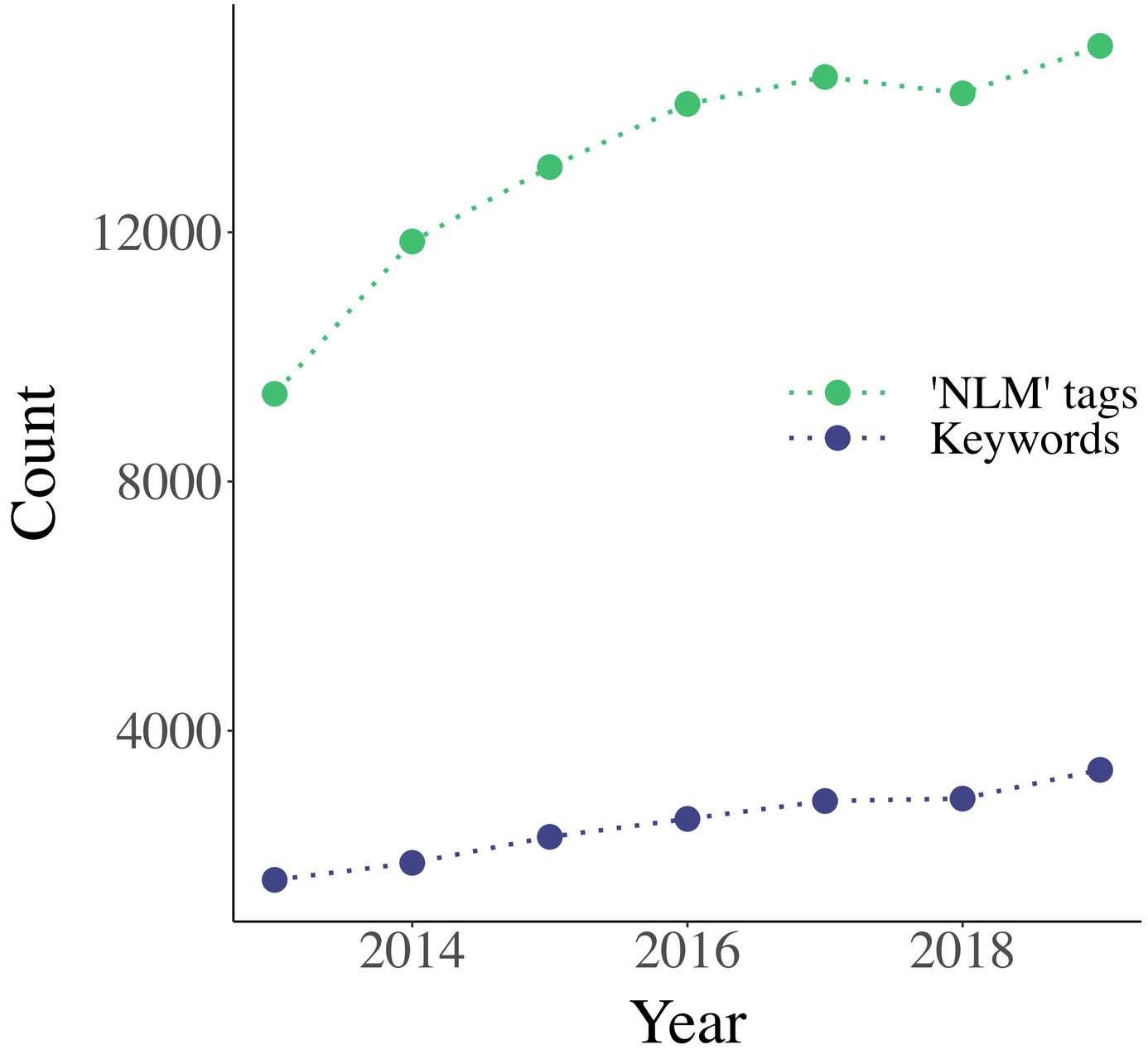
- ~ 11 percent increase in U.S.-based research
- ~ 60 percent increase in Europe-based research
- ~ 225 percent increase in China-based research



Growth explained by changes in *content*

- 50-120 percent increase in the number of meta-analyses and literature reviews

[~ “geometric increase” in meta-analyses documented in Ioannidis et al. 2013]



Takeaways

Language Models for Data Construction (generally)

- Task-specific language models allow researchers to approximate the quality of frontier LLMs, at a fraction (here: 3%!) of the cost
    - The performance of bespoke models depends on the *quantity* and *quality* of labels
    - For our binary classification task:
        - iterative refinement of prompts + model distillation kept false positive *and* false negative rates below 5%.

Trends in Clinical Trial Production (specifically)

- Since ~1990, concerns about the productivity of the pharmaceutical industry have shaped policy [see Cockburn 2004, 2006 for a review of the evidence]
  - (on drug pricing, on the structure of federal subsidies for R&D, on regulatory standards for new medicines . . .)

Key evidence:

- # new molecular entities approved by FDA *constant?* [**yes**]
- dollars spent on pharma. R&D *increasing?* [Sertkaya et al. 2024 suggests **no**]
- # of clinical trials *increasing?* [our data suggests **no**]

- *Refinement of a classification problem* suggests a very different conclusion about the productivity of this industry